

Corrigé du chapitre 9

Comprendre les données sociales

Mark Paradis, Ph. D.

2 septembre 2026

Voici le corrigé détaillé des exercices de *fin de chapitre* du chapitre 9 de *Comprendre les données sociales*. Le corrigé des exercices situés à l'intérieur de chaque section se trouve dans le livre lui-même, immédiatement après les exercices auxquels il répond.

Faites l'exercice avant de lire le corrigé. Un corrigé lu en premier n'est qu'un texte de plus ; un corrigé lu en second est une vérification de votre raisonnement, et seul le second enseigne quelque chose.

Le corrigé complet, pour tous les chapitres, se trouve à <https://www.markpara-dispolitics.com/books/comprendre-les-donnees-sociales/online/solutions/>.

Chapitre 9 La distribution normale

Corrigé de l'exercice 9.E1. (a) Le croquis devrait être une cloche symétrique centrée sur 250, avec des marques à :

$$\mu \pm 1\sigma : 210 \text{ et } 290$$

$$\mu \pm 2\sigma : 170 \text{ et } 330$$

$$\mu \pm 3\sigma : 130 \text{ et } 370$$

(b)

$$z = \frac{310 - 250}{40} = \frac{60}{40} = 1,50$$

Elle se situe à un écart type et demi au-dessus de la moyenne.

À l'aide de la règle empirique : 68 pour cent se situent à moins d'un écart type, donc 84 pour cent tombent sous $z = 1$; et 95 pour cent se situent à moins de deux, donc 97,5 pour cent tombent sous $z = 2$. Une cote z de 1,50 se situe entre les deux, donc elle est quelque part dans les quatre-vingt-dix.

La table donne le chiffre exact : $P(Z \leq 1,50) = 0,9332$, soit le 93^e centile.

(c) Le 15^e centile est une aire de gauche de 0,1500. En cherchant dans le corps de la table, l'entrée la plus proche est 0,1492 à $z = -1,04$.

$$x = \mu + z\sigma = 250 + (-1,04)(40) = 250 - 41,6 = 208,4$$

Le seuil est d'environ 208.

Attention au signe. Les élèves sous le 15^e centile sont sous la moyenne, donc la cote z doit être négative et le score obtenu inférieur à 250. Une réponse supérieure à 250 a perdu le signe en chemin.

(d) Les 5 pour cent du haut laissent 95 pour cent en dessous, donc $z = 1,645$.

$$x = 250 + 1,645(40) = 250 + 65,8 = 315,8$$

Le score minimal est d'environ 316.

(e) Standardiser les deux bornes.

$$z_1 = \frac{220 - 250}{40} = -0,75 \quad z_2 = \frac{290 - 250}{40} = 1,00$$

$$P(Z \leq 1,00) = 0,8413 \quad P(Z \leq -0,75) = 0,2266$$

$$0,8413 - 0,2266 = 0,6147$$

Environ 61,5 pour cent des élèves obtiennent entre 220 et 290.

Corrigé de l'exercice 9.E2. (a) *Cours A* :

$$z = \frac{80 - 68}{12} = \frac{12}{12} = 1,00$$

Cours B :

$$z = \frac{80 - 74}{6} = \frac{6}{6} = 1,00$$

Les cotes z sont *identiques*. Les deux étudiants se situent exactement à un écart type au-dessus de la moyenne de leur propre cours, de sorte qu'aucun n'a mieux réussi par rapport à ses camarades — la différence en écarts types est nulle.

Il vaut la peine de le remarquer avant le (d), car c'est le point autour duquel l'exercice est bâti.

(b) Les deux proportions sont les mêmes, pour la même raison.

$$P(Z \leq 1,00) = 0,8413 \quad P(Z > 1,00) = 1 - 0,8413 = 0,1587$$

Environ 15,9 pour cent des étudiants ont obtenu plus de 80 dans chaque cours.

(c) Les 20 pour cent du haut laissent 80 pour cent en dessous. En cherchant 0,8000 dans la table, on trouve 0,7995 à $z = 0,84$.

Cours A :

$$x = 68 + 0,84(12) = 68 + 10,1 = 78,1$$

Cours B :

$$x = 74 + 0,84(6) = 74 + 5,0 = 79,0$$

Le seuil du tableau d'honneur est d'environ 78 au cours A et d'environ 79 au cours B.

(d) C'est *également* facile, au sens où la comparaison peut trancher : la même proportion d'étudiants — environ 16 pour cent — obtient plus de 80 dans chaque cours.

C'est une coïncidence de ces nombres-ci plutôt qu'une vérité générale. Le cours B a une moyenne plus élevée, ce qui pousse plus d'étudiants au-dessus de 80 ; mais il a aussi un écart type plus petit, ce qui en tire moins vers la queue supérieure. Ici les deux effets s'annulent exactement.

Ce que la comparaison n'établit *pas*, c'est quoi que ce soit sur le cours qui serait plus facile en un sens de fond. Les distributions décrivent comment les notes ont été attribuées, non l'exigence de la matière ni les différences entre les deux cohortes. Un cours au barème généreux et un cours aux étudiants forts peuvent produire des distributions identiques.

Nommer cette limite est la substance de la réponse, non un ajout après coup.

Corrigé de l'exercice 9.E3. (a) Les 95 pour cent du centre emploient $z = \pm 1,96$.

$$6\,800 - 1,96(2\,400) = 6\,800 - 4\,704 = 2\,096$$

$$6\,800 + 1,96(2\,400) = 6\,800 + 4\,704 = 11\,504$$

Le nombre quotidien de pas typique va d'environ 2 100 à environ 11 500.

(b)

$$z = \frac{10\,000 - 6\,800}{2\,400} = \frac{3\,200}{2\,400} = 1,33$$

$$P(Z \leq 1,33) = 0,9082 \quad P(Z > 1,33) = 1 - 0,9082 = 0,0918$$

Environ 9,2 pour cent des adultes atteignent la ligne directrice de 10 000 pas.

(c) Le 10^e centile est une aire de gauche de 0,1000. L'entrée la plus proche dans la table est 0,1003 à $z = -1,28$.

$$x = 6\,800 + (-1,28)(2\,400) = 6\,800 - 3\,072 = 3\,728$$

Environ 3 700 pas par jour.

(d) Le 25^e centile : l'entrée la plus proche de 0,2500 est 0,2514 à $z = -0,67$.

$$x = 6\,800 + (-0,67)(2\,400) = 6\,800 - 1\,608 = 5\,192$$

Environ 5 200 pas par jour.

(e) *Pour* : le nombre de pas est une mesure à peu près continue d'un comportement physique dans une population large, et de telles variables sont souvent assez symétriques autour d'une valeur centrale. Aucun mécanisme évident ne produit une très longue queue dans un sens ou dans l'autre.

Contre : la variable est bornée à zéro par le bas, et la moyenne ne se situe qu'à 2,8 écarts types au-dessus — le modèle attribue donc à environ 0,2 pour cent des adultes un nombre de pas négatif, ce que personne ne peut avoir, et bien davantage à des nombres assez proches de zéro pour être à peine possibles chez un adulte qui marche. Il y a aussi de bonnes raisons d'attendre une asymétrie à droite, puisqu'une minorité de gens marchent ou courent bien plus que la moyenne alors que personne ne peut marcher moins que rien.

Le modèle est probablement adéquat au milieu de la distribution et le moins fiable exactement dans la queue inférieure sur laquelle portaient les (c) et (d).

Corrigé de l'exercice 9.E4. (a) Chacune à son tour.

$M = 0,53$, $\hat{ET} = 0,18$: la moyenne se situe près du milieu de l'intervalle 0–1, et l'écart type est modéré par rapport à elle — un coefficient de variation d'environ un tiers.

Asymétrie = $-0,05$: très légère asymétrie à gauche. Selon les conventions de la

§8.8.1, une valeur entre $-0,5$ et $+0,5$ compte comme approximativement symétrique, donc celle-ci est négligeable.

Aplatissement = $-0,45$: queues légèrement légères — la distribution produit moins de valeurs extrêmes qu'une courbe normale n'en produirait. Un aplatissement excédentaire négatif signifie plus plat que la normale, et une valeur de cette taille est un écart modeste.

Ensemble, ces chiffres décrivent une distribution presque symétrique et au sommet légèrement aplati.

(b) Avec $n = 340$, le test a une puissance raisonnable, de sorte qu'un résultat non significatif pèse un certain poids — contrairement à $n = 25$, où le test ne détecterait à peu près rien.

$p = 0,30$ signifie que le test n'a trouvé aucune preuve contre la normalité. Notez la formulation prudente : ce n'est pas une preuve que la distribution *est* normale, puisque ne pas rejeter une hypothèse nulle n'équivaut pas à l'accepter (§3.3.3).

Notez aussi qu'à $n = 340$ un écart important aurait très probablement été détecté, de sorte que le résultat non significatif est plus informatif ici qu'il ne le serait dans un petit échantillon.

(c) Oui, le modèle normal est raisonnable. Les chiffres d'asymétrie et d'aplatissement sont tous deux petits, le test formel ne trouve aucun écart à une taille d'échantillon de puissance convenable, et rien dans les statistiques descriptives ne suggère un problème.

(d) Un *diagramme quantile-quantile*, avant tout. Les statistiques sommaires décrivent la forme d'ensemble de la distribution mais ne peuvent pas montrer où se situe un écart, et un diagramme quantile-quantile confirmerait le portrait ou révélerait quelque chose que les sommaires ont manqué — la bimodalité en particulier, que l'asymétrie et l'aplatissement peuvent tous deux ne pas enregistrer.

Un histogramme servirait le même but. La médiane serait utile aussi : avec une moyenne rapportée de $0,53$ et une légère asymétrie négative, la médiane devrait se situer un peu au-dessus, et la vérifier ne coûte rien.

(e) Cela compte, et c'est l'argument le plus fort contre le modèle.

Le score est borné à 0 et 1 , alors que la distribution normale s'étend à l'infini dans les deux directions. Avec $M = 0,53$ et $\acute{E}T = 0,18$, zéro se situe à $2,94$ écarts types sous la moyenne et 1 à $2,61$ au-dessus, de sorte que le modèle attribue environ $0,2$ pour cent de sa probabilité sous zéro et environ $0,5$ pour cent au-dessus de un — des valeurs que la variable ne peut pas prendre.

Notez laquelle des deux est la plus grande. La borne supérieure est la plus rapprochée en écarts types, donc la fuite au-dessus de un est la plus importante des deux, et un croquis de la courbe ajustée la montrerait débordant par le haut de l'échelle en premier.

En pratique, 0,6 pour cent au total est une fuite assez petite pour être ignorée dans la plupart des usages. Mais elle explique l'aplatissement négatif : une variable bornée ne peut pas produire les longues queues qu'une distribution normale attend, donc ses queues sont nécessairement légères. Le bornage et le chiffre d'aplatissement sont le même fait vu deux fois.

Corrigé de l'exercice 9.E5. 1. La table donne des aires de gauche, et 0,9821 est $P(Z \leq 2,10)$. La question demandait la queue de droite.

$$P(Z > 2,10) = 1 - 0,9821 = 0,0179$$

Un croquis l'aurait attrapé : l'aire au-dessus de $z = 2,10$ est un mince filet, manifestement pas 98 pour cent de la distribution.

2. Standardiser change les unités, non la forme. Le revenu se standardise en une distribution de moyenne 0 et d'écart type 1 qui est *toujours asymétrique à droite*. Les cotes z sont correctes comme descriptions de position ; ce qu'elles n'autorisent pas, c'est de lire des probabilités dans la table normale.

3. Un intervalle de référence contenant les 95 pour cent du centre exclut 5 pour cent de la population *en santé* par construction — 2,5 pour cent à chaque bout. En sortir signifie être statistiquement inhabituel, non être malade. Une personne en santé sur vingt tombe à l'extérieur, et un seul résultat hors intervalle appelle normalement une reprise de l'analyse plutôt qu'un traitement.

4. Une moyenne de 4 200 contre une médiane de 2 900 est un écart très important, qui indique une forte asymétrie à droite. La règle empirique est une propriété de la distribution normale et ne se transporte pas aux données asymétriques — les proportions réelles à moins d'un et de deux écarts types différeront de 68 et 95 pour cent, et les intervalles ne seront pas centrés là où se trouve le gros des données.

5. À $n = 8\,000$, le test de Shapiro–Wilk rejettera à peu près n'importe quelle distribution réelle, parce qu'aucune n'est exactement normale et que le test a une puissance énorme à cette taille d'échantillon. Le résultat dit qu'un écart existe ; il ne dit rien sur la question de savoir si l'écart est assez grand pour compter.

Le chercheur aurait dû examiner un diagramme quantile-quantile pour juger de la taille et de l'emplacement de l'écart avant de renoncer à quoi que ce soit.

6. Le nombre de frères et sœurs est un *dénombrement* : discret, borné à zéro par le bas et asymétrique à droite. Un modèle continu symétrique réparti de la probabilité sur des valeurs impossibles — 2,1 frères et sœurs, cela n'existe chez personne — et avec $\mu = 2,1$ et $\sigma = 1,6$, le modèle place environ 9 pour cent des répondants sous zéro frère ou sœur.

Un tableau de fréquences des dénombrements réels serait à la fois plus informatif et moins de travail.

Corrigé de l'exercice 9.E6. Vos réponses dépendent des variables que vous choisissez, mais la forme du raisonnement, non.

(a) Faites les prédictions avant de tracer. C'est tout l'intérêt de cette partie, et inverser l'ordre la transforme discrètement en exercice de description — vous vous retrouverez à expliquer ce que vous avez déjà vu.

Choix raisonnables : `age` ou `policy_support` pour la variable approximativement normale, `income` pour celle qui ne l'est pas. Le raisonnement devrait invoquer les caractéristiques de fond de la §9.6.5 — mesure continue dans un groupe homogène contre argent, dénombrements ou durées.

(b) Les histogrammes et les diagrammes quantile-quantile devraient montrer ce que montrent les figures 9.5 à 9.15 du chapitre : un accord étroit pour la variable bien élevée, et pour le revenu un histogramme visiblement asymétrique avec une courbe ajustée qui descend sous zéro, plus un diagramme quantile-quantile qui s'incurve vers le haut à droite.

(c) Pour la variable approximativement normale, la moyenne et la médiane seront proches et le coefficient d'asymétrie petit. Pour le revenu, la moyenne dépasse nettement la médiane et le second coefficient de Pearson sort à 0,79 — bien au-dessus du 0,5 que la §8.8.1 traite comme la limite de la symétrie approximative.

(d) Ils divergeront, et pas de peu. Shapiro–Wilk rejette *toutes* les variables candidates de `socsurvey` — âge, appui aux politiques, idéologie et revenu confondus. La plus grande valeur p parmi elles est 0.0000014, ce qui n'est pas un résultat limite.

Ce n'est pas une découverte sur `socsurvey`. C'est ce que la §9.6.4 prédit à toute taille d'échantillon raisonnable : aucune variable réelle n'est exactement normale, et le test répond à la question « est-ce exactement normal ? »

Là où ils divergent, fiez-vous au graphique — pour la raison donnée à la §9.6.4. Le test répond à « est-ce exactement normal ? », dont la réponse n'est jamais oui ; le graphique répond à « l'écart est-il assez grand pour compter ? », qui est la question qui gouverne la suite.

(e) Rapportez le chiffre que vous obtenez réellement, et résistez à l'envie de l'arrondir vers 68. Ce que vous trouvez dépend de la variable plus que vous ne le penseriez :

Variable	À moins d'un ET
<code>age</code>	66,2 %
<code>policy_support</code>	67,7 %
<code>ideology</code>	76,1 %
<code>income</code>	74,0 %

L'âge et l'appui aux politiques tombent près de 68, ce à quoi la règle empirique fait s'attendre.

L'idéologie non, à 76,1 pour cent — et c'est la variable que ce chapitre emploie

comme exemple de bon ajustement normal (figure 9.5). Il n'y a pas de contradiction. L'idéologie est une échelle discrète en onze points avec beaucoup de masse entassée au centre, de sorte qu'une cloche symétrique décrit bien sa *forme* pendant que sa proportion à moins d'un écart type sort élevée. La règle empirique est une propriété d'une courbe normale continue, et une échelle discrète et bosselée peut s'en écarter dans un sens ou dans l'autre sans que le modèle normal soit une mauvaise description.

Le revenu sort élevé lui aussi, à 74,0 pour cent, pour une raison entièrement différente : la moyenne a été tirée dans la queue, de sorte que l'intervalle autour d'elle couvre davantage de la région inférieure dense qu'une distribution symétrique n'y en placerait.

Deux variables, toutes deux bien loin de 68 pour cent, pour des raisons opposées — ce qui est une meilleure leçon qu'une confirmation bien rangée.